12 BS

ARPA ORDER NO. 2930
NR 154-374

ISI/RR-77-54
January 1977

William C. Mann
James H. Carlisle
James A. Moore
James A. Levin

[redacted] An Assessment of Reliability of Dialogue-Annotation Instructions

INFORMATION SCIENCES INSTITUTE
UNIVERSITY OF SOUTHERN CALIFORNIA ISI
4676 Admiralty Way/ Marina del Rey/ California 90291
(213) 822-1511

Preparation of this paper was supported by the Office of Naval Research, Personnel and Training Research Programs, Code 458, under Contract N00014-75-C-0710, NR 154-374, under terms of ARPA Order Number 2930.

The views and conclusions contained in this document are those of the author(s) and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Office of Naval Research, the Defense Advanced Research Projects Agency, or the U.S. Government.

This document is approved for public release and sale; distribution is unlimited.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

14 1. REPORT NUMBER: ISI/RR-77-54

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

6 4. TITLE (and Subtitle): AN ASSESSMENT OF RELIABILITY OF DIALOGUE ANNOTATION INSTRUCTIONS,

9 5. TYPE OF REPORT & PERIOD COVERED: Research rept.,

6. PERFORMING ORG. REPORT NUMBER: -------

10 7. AUTHOR(s): William C. Mann, James H. Carlisle, James A. Moore, James A. Levin

15 8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0710, ARPA Order-2930

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
USC/Information Sciences Institute
4676 Admiralty Way
Marina del Rey, California 90291

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:
61153N RR042-06-01
RR042-06 NR154-374

11. CONTROLLING OFFICE NAME AND ADDRESS:
Cybernetics Technology Office
Advanced Research Projects Agency
1400 Wilson Blvd., Arlington, VA 22209

11 12. REPORT DATE: Jan[redacted] [redacted]77

13. NUMBER OF PAGES: 64

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):
Personnel and Training Research Programs
Office of Naval Research - Code 458
800 No. Quincy St.
Arlington, VA 22217

15. SECURITY CLASS. (of this report): UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

12 63p.

16 17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report): RR04206

17 18. SUPPLEMENTARY NOTES: RR0420601

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

artificial intelligence, cognitive psychology, computer, dialogue, evaluation, linguistic, observation, reliability, research methodology, text analysis, theory.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(OVER)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

407952

YB

20. ABSTRACT

This report is part of an ongoing research effort on man-machine communication, which is engaged in transforming knowledge of how human communication works into improvements in man-machine communication of existing and planned computer systems. This research has developed some methods for finding certain kinds of recurring features in transcripts of human communication. These methods involve having a trained person, called an Observer, annotate the transcript in a prescribed way. One of the issues in evaluating this methodology is the potential reliability of the Observer's work.

This report describes a test of Observer reliability. It was necessary to design a special kind of test, including some novel scoring methods. The test was performed using the developers of the instructions as Observers.

The test showed that very high Observer reliability could be achieved. This indicates that the observation methods are capable of deriving information which reflects widely shared perceptions about communication, and which is therefore the right kind of data for developing human communication theory. It is a confirmation of the appropriateness and potential effectiveness of using this kind of observations in the dialogue-modeling methodology of which they are a part. It is also of particular interest as an approach to study of human communication based on text, since content-related text-annotation methods have a reputation of low reliability.

ARPA ORDER NO. 2930
NR 154-374

ISI/RR-77-54
January 1977

William C. Mann
James H. Carlisle
James A. Moore
James A. Levin

An Assessment of Reliability of Dialogue-Annotation Instructions

INFORMATION SCIENCES INSTITUTE
UNIVERSITY OF SOUTHERN CALIFORNIA
4676 Admiralty Way/ Marina del Rey/ California 90291
(213) 822-1511

Preparation of this paper was supported by the Office of Naval Research, Personnel and Training Research Programs, Code 458, under Contract N00014-75-C-0710, NR 154-374, under terms of ARPA Order Number 2930.

The views and conclusions contained in this document are those of the author(s) and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Office of Naval Research, the Defense Advanced Research Projects Agency, or the U.S. Government.

This document is approved for public release and sale; distribution is unlimited.

LIST OF TABLES AND FIGURES

TABLES

FIGURES

AN ASSESSMENT OF RELIABILITY OF DIALOGUE ANNOTATION INSTRUCTIONS

ABSTRACT

This report is part of an ongoing research effort on man-machine communication, which is engaged in transforming knowledge of how human communication works into improvements in the man-machine communication of existing and planned computer systems. This research has developed some methods for finding certain kinds of recurring features in transcripts of human communication. These methods involve having a trained person, called an Observer, annotate the transcript in a prescribed way. One of the issues in evaluating this methodology is the potential reliability of the Observer's work.

This report describes a test of Observer reliablity. It was necessary to design a special kind of test, including some novel scoring methods. The test was performed using the developers of the instructions as Observers.

The test showed that very high Observer reliability could be achieved. This indicates that the observation methods are capable of deriving information which reflects widely shared perceptions about communication, and which is therefore the right kind of data for developing human communication theory. It is a confirmation of the appropriateness and potential effectiveness of using this kind of observations in the dialogue-modeling methodology of which they are a part. It is also of particular interest as an approach to study of human communication based on text, since content-related text-annotation methods have a reputation of low reliability.

ACKNOWLEDGMENTS

We give special acknowledgments to Rose Kattlove for her energetic work on the Similar Expressions tests below, and to Dr. Harry O'Neil of ARPA-HRRO, who made the initial suggestion to perform the tests.

I. Overview and Research Context

Following an introduction to the key problems in assessing ·eliability for systematic observational techniques, a brief description of the Dialogue Annotation Instructions (called the DAI below)(Mann, Moore, Levin, Carlisle, 1975) is presented.

In the next section, the key problems of assessing observer reliability for the DAI in particular are examined and an agreement assessment algorithm is specified in detail. Possible sources of bias and alternative algorithms are considered. Following that general presentation of the reliability assessment algorithm, a study is reported in which reliability is assessed among the four Observers across four dialogues. Results and discussion are presented for each annotation category of the DAI. The final secticn, contains summaries of the reliability assessment algorithms and the results of the study.

II. Reliability in Systematic Observation Techniques

The attainment of reliability has long been a difficult task in the development of systematic observational techniques. Typically, reliability is defined in terms of degree of agreement between independent Observers. Heyns and Lippitt (1954), in their review of observational techniques for The Handbook of Social Psychology implore that

> "Because of the difficulties involved in obtaining satisfactory reliability and the responsiveness of reliability scores to training, it is virtually mandatory that reliability checks be run prior to securing the research data." (p. 397)

High reliability is particularly difficult to obtain when much inference is required of the Observer. Unless the observation judgments are trivial, differences among Observers in interpretation and execution of the rules will result in significantly less than perfect reliability. Reliability can be increased by arbitrarily determining the units to be classified (e.g., time intervals) or by providing a limited number of mutually-exclusive categories. In the observation task with which this paper is concerned, the observer is responsible for identification as well as classification of complex units on a variety of dimensions. It is, therefore, to be expected that high reliability will be difficult to attain.

Many techniques increase reliability of scoring by comparing only the total or relative frequencies of occurrence of different types of unit. However, reliability must be computed on judgments at the level for which observations are to be used in analysis. Bales (1951), for example, has reported reliability scores ranging from .75 to .95 for agreement as to the number of acts which fall into each category for each individual in a group meeting. Disagreement with respect to single acts tend to average out across categories over a large sample of data. When the theoretical analysis is of individual units of behavior (such as speech utterances) rather than relative frequency of each type of units, then observer reliability must be computed with respect to the annotation of individual units.

The assessment of reliability for systematic observational techniques involving a high level of observer inference and observer identification of units must necessarily deal with dependencies. Some judgments by Observers may be dependent upon previous observations. If event E1 would not be annotated at all unless event E had been identified and annotated, then the computation of reliability for annotating E1 should take into consideration its prerequisite. Disagreement as to the annotation of E should lower the reliability score, as should disagreement on E1 among those Observers who agreed with respect to E. This leads to the notions of nesting and levels of annotation for which reliability can be computed. However, it is desirable to account for prerequisites so that an overall score of reliability could be determined.

What is the importance of testing Observer reliability? Only that in supporting certain kinds of scientific claims, reliability is a premise. There are many feasible uses of the sort of observation method that we are studying here, with correspondingly many kinds of claims, which we will not explore here.

In the methodology of this project, the scientifically significant results are particular ***processes***, specifiable as computer algorithms, which embody some knowledge of human communication. These processes are fragments of dialogue models, which are partial simulations of actual human dialogues. The scientific claims that we make refer to these processes. (For example, consider a process which was able to detect an appeal for help when it occurred in a dialogue.)

An important form of claim about a process P is as follows:

P represents a ***widely-shared*** interpretive regularity of human communication.

(There are 5 technical terms in this claim form: P, represents, widely-shared, interpretive regularity, and human communication. Some of these are further specified below.)

An interpretive regularity is an attribute of an individual person. It is some demonstrable pattern of his responding to the kind of information which comes to him expressed in symbols. It can be expressed as a set of conditions and a set of consequences. If we have an Observer O annotate a dialogue, his assertions about the implicit structure of the dialogue are evidence of his own interpretive regularities. We might claim that a process P represents a particular interpretive regularity of O, and, to support that claim, we might obtain evidence about the correspondence between the invocation conditions and execution consequences of P, on one hand, and the dialogue conditions and observation assertions of O on the other. (If we are very successful in this general enterprise, we may build a comprehensive model, composed of many processes, of O's interpretive regularities.) We then have an evidenced claim that :

P represents an interpretive regularity of O's human communication.

At this point, Observer reliability becomes relevant. If we have evidence that those interpretive regularities of O which are represented in his observations are also held by many other potential observers, then we can make the additional claim that

P represents a ***widely-shared*** interpretive regularity of human communication

which is the kind of claim we wanted to make. High Observer reliability is evidence that interpretive regularities are ***widely-shared.***

This paper describes and demonstrates a methodology for assessing reliability for high-inferential nested systematic observation techniques. This methodology should be applicable to a variety of complex observation techniques, such as protocol analysis (Newell, 1966; Newell & Simon, 1972). For an alternative approach to repeatable text analysis, see Waterman (1973). We are, in this paper, primarily concerned with definition and use of the methodology to assess reliability of the Dialogue Annotation Instructions.

III. The Dialogue Annotation Instructions

The DAI were developed at the USC/Information Sciences Institute to facilitate study of particular aspects of the human ability to communicate. The annotation of actual dialogues with respect to phenomena such as:

requests
repeated references
topic structure
expressions of comprehension
similar expressions

provides data to be used in the development of theories and computer models to account for the understanding of human dialogue. The goal of the overall research effort is "to significantly expand and diversify the capabilities of the computer interfaces that people use. The approach is to first design computer processes that can assimilate particular aspects of dialogue between people, then to transfer these processes into man-machine communication" (Mann, et. al., 1975.) This overall research effort has been described elsewhere (Mann, 1975).

A brief description of each of the annotation categories is provided below to characterize the need for the reliability assessment algorithm and to facilitate interpretation of the results of the study reported later in this paper. The reader is referred to the actual Dialogue Annotation Instructions (DAI) for a complete description of the annotation categories (Mann, et. al., 1975.) In reading the summary category descriptions below, note the high degree of reliance placed on Observers to identify categorical events and to qualify them in detail.

Observers are given the DAI to study and after several practice annotation and discussion sessions are presented with a transcript of an actual dialogue to be annotated. A fresh copy of the transcript is used for each category. Observers are asked to note only those instances which they regard as *clearly* corresponding to the instructions. Special conventions are introduced for annotating segments of text, and for labeling these segments. The following categories are then annotated, one at a time.

Requests

The observer is asked to locate all places in the dialogue where a speaker communicates to the hearer a specific expectation he has of the hearer's future behavior. Based on the immediacy of the expected behavior, whether it is verbal or non-verbal, if the request is not intended to be taken literally, and if what is requested is the absence of certain behavior, the observer is asked to characterize the Request as one of five specified types (Question, Order, Directive, Rhetorical or Prohibitive).

For each such Request he notes, the observer is also to characterize the response as compliant or not. Next, in most cases, he is asked to choose which type of response compliance (from a given set of types) best describes the actual response. Finally, he is to judge whether or not the request was ever complied with. Compliance is defined as providing (or beginning to provide) the requested behavior.

Repeated Reference

On the assumption that asking the observer to encode the object/concept target of each referring expression was hopelessly intractable, we opted instead to have the observer note whenever two expressions which occurred in the dialogue, were used to refer to the same thing. Special instructions cover the cases of reference to the participants themselves, and references to segments of the dialogue, as uninterpreted text.

The initial version of the instructions contains a part dealing with references to elements (and subsets) of sets. Our early experiences with these showed them to be difficult to perform and interpret, so they were dropped*.

Topic Structure

The observer is instructed to note the points in the dialogue where each participant initiates or accepts a topic as well as the points where each appears to close or abandon the topic. Whenever the observer judges that a participant first exhibits dialogue relevant to a topic, that point in the dialogue is to be annotated with a "begin" mark and a short name for the corresponding topic. Similarly, for the place where the same participant last seems to be influenced by the previously-opened topic, that point is to be noted with an "end" mark and the same label that was invented for the corresponding "begin". In the case for which the observer judges that the participants are sharing a topic, the same name is used in both cases. When the points of topic beginnings and endings are less distinct, there is a notation for indicating this. Finally, the observer is asked to name all topics which were apparently already begun before the dialogue segment being examined, as well as those which seem to continue beyond the segment's end.

----

*This is one of two instances in which a small portion of the instructions was not used because we had already decided to eliminate that portion in future versions of the instructions.

Expression of Comprehension

The Observer is asked to locate all places in the dialogue where one participant indicates, in some way, his degree of comprehension of some aspect of his partner's prior conversation. He may indicate that he does understand (Positive Comprehension) or that he does not (Negative Comprehension). ,ie may indicate that his comprehension (or lack, thereof) is partial (Selective Positive Comprehension, Selective Negative Comprehension). Finally, the Observer is to judge whether the utterance which exhibited this degree of comprehension had that function as its sole purpose (Positive Primariness) or not (Negative Primariness). (Annotation of the strength of comprehension, using labels P1, P2, N1, N2, was not performed.)

Similar Expressions Out of Context

There are five steps to this type of annotation. First, a dialogue is divided by the experimenter into units, each having approximately the "completeness" of a simple English sentence. Second, all such units from several dialogues are mixed together so that order and source are obscured. Third, a native English speaker, who is not one of the Observers, generates for each unit (out of context) three "similar expressions."

Fourth, Observers are presented with the similar expressions generated out of context, arranged into groups. One unit in each group is designated as the standard unit (original from the dialogue); the others are comparison units. Observers are asked to score each comparison unit as to acceptability as a substitute for the standard in some ordinary circumstances. Fifth, Observers are given a complete transcript with units numbered and, for each unit, the set of those expressions which were judged similar in the preceding step. These are then evaluated *in context* for acceptability. The acceptability annotations generated in steps four and five are the items which we evaluate for Observer agreement*.

----

* An additional category of annotation, Correction Events, is defined in the DAI. We chose not to test reliability in this category for several reasons. It would have required a separate corpus of dialogue, since correction events are low-probability events. It would have been the most complex and time-consuming category. The definition style for Correction Events is very much like that for other categories (particularly Requests) so that it would tend to stand or fall with the other categories, being therefore somewhat redundant. We expect that Observer reliability for Correction Events will be tested in future tests.

IV. The Methodology for Reliability Assessment

A. Design Issues for a Reliability Assessment Method

The nature of an appropriate test of Observer reliability is strongly shaped by the details of the Observer's task and by perceptions about what kinds of agreement between Observers are significant. For the DAI the methods used to compare annotations by different Observers must be selected with care, particularly with respect to the following issues:

1. The method should yield enough information about details of observation to be useful for improving the Observer's Instructions.

2. Differences between essentially arbitrary parts of the annotations must be treated as insignificant. (Example: the arbitrary labels chosen by observers for particular units.)

3. The method should not be excessively sensitive to the bulk of material being judged. (Example: recognizing a long question should be counted equal with recognizing a short one.)

4. The method of judging agreement must be capable of measuring an uncontrolled number of judgments, since the Observer is free to select where he will annotate.

5. The comparison method must be simple and homogeneous enough to be readily understood.

The DAI yield a rich variety of annotations, many of which are assertions about ranges of text. In order to make a simple, uniform algorithm applicable, all range-like annotations (except on Topic Structure) are collapsed into single word "events" as part of the agreement assessment process. These collapsing transformations were defined before performing the test. They are defined in full in *Section V.A* below.

The hierarchical nature of the DAI and the event coding scheme permit scoring of Observer reliability at various levels of specificity, so that unreliability in certain kinds of judgments is not masked by overall high reliability. For example, consider the annotation of Requests: identification of the type of a Request, classification of the partner's immediate response and judgment of whether and how the partner eventually responded to the request - all these are distinguished and judged separately for Observer agreement.

This is therefore a sensitive probe into the strengths and weaknesses of both the annotation instructions and the Observers. We expect to be guided in part by these results in preparing any future versions of the observation methods.

B. The Agreement Assessment Algorithm

1. *Event Collapsing*

For each of the observation categories, a set of annotations is transformed into a sequence of events which appear in text order. The events are indexed to the text, so that it is unambiguous whether an event in each of two observation sets occurred at the same place.

Observation events have properties, almost all of which are direct transcriptions of annotation marks which the Observer is instructed to use. All of the possible event properties are known in advance and drawn from small finite sets. (Even though Observers are allowed to comment, no free-prose annotations are examined for agreement assessment.)

The algorithm must measure agreement of sequences of propertied events. The method is first explained by example for the dominant simple case of event identification (which will be referred to hereafter as Level One agreement), then by example for the more complex dependent-annotation case (Level Two agreement), and finally by a pseudo-program for the general case.

2. *Agreement on Event Identification (Level One)*

An example of annotations by three Observers for a short 14 word section of text is shown in Figure 1 below. The column at the left corresponds to the word numbers. The labels used by Observers in their annotations are F,B,C,D, and E. Considering each annotation, one at a time, the numbers of actual [A] and possible [P] agreements are shown in Figure 2, where the fractions are A/P. It can be seen that at word 2, all Observers agreed that an event of type F occurred, yielding A=6 and P=6. There were 6 out of 6 possible agreements, likewise at word 5. At word 9, only two Observers asserted that an event of type C occurred. Each of these two observations had 1 out of a possible 2 agreements, giving 2 out of 4 possible agreements

Word	Observer 1	Observer 2	Observer 3
1			
2	F	F	F
3			
4			
5	B	B	B
6			
7			
8			
9	C	C	C
10			
11			
12			
13	D	E	
14			
15			

Figure 1. An Example of Pairwise Comparison

for that event. At word 13, there were no agreements, but each of the 2 observations made cculd have had 2 agreements (thus 0 out of 4). The reliability ratio for this example is computed as follows

at word	actual agreements	possibie agreements
2	6	6
5	6	6
9	2	4
13	0	4
	14	20

giving 14 out of 20 possible agreements for a reliability of .70.

Word	Observer 1	Observer 2	Observer 3	Observer 4
1				
2				
3	A	A	A	A
4				
5	B	B	B	
6				
7	B1	B1	B1	
8				
9	B2	B2		
10				
11				
12		C	C	C
13				
14		D	E	E
15				
16		C1	C1	C1
17				

Figure 2. An Example of Pairwise Comparison with Prerequisites

3. *Agreement on Event-Dependent Annotations (Level Two)*

Figure 3 shows a similar stream of encodings, this time with 4 Observers. There is a second kind of events shown, marked with B1, B2 and C1. These are events that can only be asserted by an observer provided some prerequisite observations have been made first. Here the intent is that events B1 and B2 have B as a prerequisite, and event C1 has C as a prerequisite. We shall use the term "Level Two" to refer to those observations for which prerequisite observations exist.*

For the example in Figure 3, the observation events which had no prerequisites (ie, Level One) are scored as in the previous example,

at word	actual agreements	possible agreements
3	12	12
5	12	12
12	6	9
14	2	9
	26	39

giving 26 out of 39 possible agreements for a reliability ration of .66. The events which did have prerequisites take that fact into consideration in calculating the possible number of agreements. Thus, for Figure 3, the Level Two reliability is computed from

at word	actual agreements	possible agreements
7	6	6
9	2	6
16	6	6
	14	18

giving 14 our of 18 possible agreements for a ratio of .78.

----

* The term "level two" is somewhat misleading in that it refers to any subordinated level for which preconditions to observations are taken into account.

Word	Observer 1	Observer 2	Observer 3
1			
2			
3	A	A	A
4			
5	B	B	B
6			
7	B1	B1	B1
8			
9	B2	B2	
10			
11			
12	C	C	C
13			
14	D	E	E
15			
16	C1	C1	C1
17			

Figure 3. An Example of Pairwise Comparison With Prequisities

The summary over all events for this example gives 40 out of 57 possible agreements for an aggregate reliability ratio of .70. Note that the observation of events with prerequisites can be highly reliable even if observation of the prerequisite events has a low reliability.

4. *Combining Reliability Scores*

The computational methods for combining reliability scores are fairly simple and straightforward. Within observational categories, they are all designed on a one-observation : one-vote basis, with aggregation done by a method which, in effect, treats the various subgroups of observations in the category as being of the same kind in the aggregate.

Let P[i] be the number of possible agreements for a single annotation i, and A[i] the number of agreements actually achieved. Then the observational reliability for annotation i is the ratio

$$R[i] = \frac{A[i]}{P[i]} \qquad (1)$$

The reliability for any set of annotations is computed by (summing separately the A[i] and P[i] which gives) the ratio

$$R[i] = \frac{\sum A[i]}{\sum P[i]} \qquad (2)$$

If the k th kind of observation is an aggregate of the i th and j th kinds, then its reliability is computed as

$$R[k] = \frac{\sum A[i] + \sum A[j]}{\sum P[i] + \sum P[j]} \qquad (3)$$

and generally to aggregate m independent kinds of observations into a single reliability measure, the ratio of sums

$$R[m] = \frac{\overset{m}{\sum_{i=1}^{m}} A[i]}{\sum_{i=1}^{m} P[i]} \qquad (4)$$

This same formula (4) is applied for all within category computations.

Notice that the reliability assessment formula is the same for all categories, in spite of their diversity and that it is the same for minor subcategories or single annotations.

For the overall reliability score, we compute an average reliability across categories in the conventional way. The aggregation formula above is not used across categories because we wish to avoid domination of the overall reliability by the one or two categories that contain very large proportions of the observations.

The particular reliabilities calculated and the identification of the subcomponents of aggregate reliabilities are described in *Section VII*.

5. *Sources of Possible Bias*

There are numerous factors which tend to systematically increase or decrease reliability scores when analyzing the same annotation data with different reliability assessment algorithms. The methodology described above is *conservative* with respect to many of these factors (*viz*, all we could think of).

The scoring method relies on segment collapsing in order to make straightforward the use of a uniform reliability computation method. The collapsing method removes many "irrelevant" differences between comparable observations, but we find that it also retains some rather unfortunate differences which do not reflect genuine differences in Observers' perception. Often, moving a region boundary by one or two words would have converted a disagreeing pair of observations into an agreeing pair.

This test should be interpreted as measuring degrees of "co-assertion" rather than "agreement of opinion" among Observers. To score an observation agreement in this test, the Observers must independently decide that the phenomenon being coded is CLEARLY

PRESENT at a particular point. An Observer might be doubtful or neglect to make some annotation which he would have regarded as correct, were he asked explicitly. Thus he fails to create an agreeing annotation, without any actual disagreement of view. Thus it is quite possible to have low reliability scores and yet have the observations faithfully reflect widely-shared perceptions about communication. Qualitatively, we are quite sure that this kind of difference contributes importantly to the level of unreliability found. On the other hand, it is implausible that high reliability would be achieved on this test without widely-shared similarities of perception of the communication.

Similarly, the test seems relatively vulnerable to differences among Observers in "sensitivity" or confidence, which result in different rates of annotation by different Observers. One observer may view part of a dialogue as having a single topic, where another sees a topic with five distinct subtopics. Their annotations could thus diverge widely even though they had the same communication interpretation of the text. All such differences in observer sensitivity tend to reduce numerical reliability. The test is also vulnerable to single idiosyncratic Observers, although this has not been a problem in fact.

Individual reliability scores were computed for Observers for several kinds of annotations and found to be uniform. (See the discussion of results in *Section VII.B on Repeated Refences.*)

A priori, it is plausible that reliability might depend on the genre of dialogue being annotated. The dialogues for this test were taken from two rather different sources. Systematic differences included:

Apollo-13	TENEX Link
Oral	Typed
Peer relation	Novice to expert relation
Parties known to each other	Strangers
Extended communication	Single complete episodes
Potentially high error cost	Low error cost

For several categories (as indicated in the specific results section below) reliability was calculated separately for each dialogue source, and no significant dependencies of reliability on dialogue source was formal.

Of course, in examining possible biases, it must be understood that this is a test of observation reliability among Observers who are deeply familiar with the method, since they are its developers. Another group of Observers might be more or less accurate, more or less conscientious, more or less aware of the nature of the judgements requested. The present Observers may also be sharing some understandings not actually written in the instructions. We expect that another group of Observers, trained for the purpose of replicating this test, would have reliability scores which were lower than those reported here by some unknown degree. As Heyns and Lippitt (1954) have pointed out, one of the best ways to maximize Observer agreement is to involve Observers in the development (or evolution) of the coding rules.

6. *Mathematical Properties of the Reliability Computation Method*

The reliability numbers which result from this test are sampling estimates of

THE PROBABILITY THAT, GIVEN A RANDOMLY SELECTED OBSERVATION BY A FIRST OBSERVER, AND A RANDOMLY SELECTED SECOND OBSERVER (from an infinite population, not depleted by removing the first observer), THE SECOND OBSERVER ASSERTS AN OBSERVATION WHICH AGREES WITH THE GIVEN OBSERVATION.

There is a downward numerical bias in our computed reliabilities relative to this interpretation, as follows: The agreement computation derives the proportion, over all observations, of other observations that agree. These other observations are necessarily by Observers other than the one producing the comparison observation. For small numbers of Observers, as in our case, this significantly biases the reliability toward smaller numbers. This bias could have been removed by an appropriate mathematical transformation, but we did not choose to do so.

So, for example, taking the case in which 50% of the population of Observers would make a particular observation, and the other 50% would not assert anything at that point, for various numbers of Observers we would have:

TABLE 1
APPARENT RELIABILITIES
FOR VARIOUS NUMBERS OF OBSERVERS
WHEN EXACTLY ONE HALF OF OBSERVERS AGREE

Number of Observers:	4	6	8	10	20	Infinite
All Observers Annotate An Event:	.17	.20	.21	.22	.24	.25
Only Half of All Observers Annotate And Those Agree:	.39	.40	.43	.44	.47	.50

This downward bias is, of course, only "relative" to other forms of reliability computation. The reliabilities computed with our pairwise agreement algorithm should not be compared directly with correlation scores or other measures of reliability. Rather, interpretation in comparisons to what sort of average behavior would be required to generate such a score. Table 2 below shows some relevant comparison points for interpreting pairwise agreement scores, and define the descriptive labels used in this report.

TABLE 2
COMPARISON POINTS FOR RELIABILITY SCORE INTERPRETATION

Reliability	A/P	Indicates an Average of	Descriptive Label Used in This Report
.00	0/12	No Observers Agree When All Four Annotate An Event	Zero Reliability
.17	2/12	Two Observers Agree When All Four Annotate An Event	Very Low Reliability
.33	2/6	Two Observers Agree When Only Two Annotate An Event	Low Reliability
.50	6/12	Three Observers Agree When All Four Annotate An Event	High Reliability
.75	9/12	Unanimous Agreement On Half Of The Events And Three Out Of Four Observers Agree On The Other Half	Very High Reliability
1.00	12/12	Unanimous Agreement On All Events Annotated	Perfect Reliability

The reliabilities which we achieved in the study reported below are much higher than could be explained by a hypothesis of random observation. We have informally estimated the random-observation reliabilities for the Level One varieties of observation for each of our major categories of observation, based on the rates of production of observations which actually occurred in this test. The estimates are in Table 3 below.

TABLE 3
ESTIMATED RELIABILITY UNDER RANDOM OBSERVATION

	Estimated Random	Comparable Actual
Requests	.24	.74
Reference	.02	.76
Topic	.34	.67
Expression of Comprehension	.64	.88
Similar Expressions	.47	.81

It is evident that random observation would not produce the levels of agreement which occurred for any category.

The reliability measure used uniformly in this test was selected in preference to correlation techniques. It fits better the conditional, observer-selected and hierarchic character of this kind of observations. However, they are nominally comparable, since one expects agreement correlations to be positive, and since the upper limit of the ranges of correlations and of our reliabilities is 1. To perform a nominal comparison between measures, one could regularize a body of data to eliminate the features that make correlation inapplicable. This would be an interesting interpretive exercise to consider including as part of a future test.

7. *Rejection of Other Algorithms for Reliability Computation*

A number of algorithms for computing reliabilities were suggested and then rejected. The reasons for rejection point up some of the properties of the method chosen.

One class of suggestions deals with scoring the various Observers against a standard "correct" set of observations. Two problems arise: Since we have the "world's most expert crew of Observers" as the observation team for this experiment, and since they are equally expert, we could not justify any particular one as the "correct" one. Even if this were done, it would not yield an independent standard. The chosen method treats the Observers as equally expert. ***Its results are identical to those which would be obtained if the sets of observations of each of the team were regarded as the "correct" standard in turn, and the results averaged.***

A second group of possible algorithms would avoid the transformation of ranges of text into observation events by scoring on a word-by-word basis. This would unfairly weight the long ranges. It would treat recognition of a long request as more significant than recognition of a short one, which seems to be directly opposite to the difficulty of the identification task. Since long phrases and short phrases can be equally valid instances of the kinds of communication phenomena under study, we prefer to weight them equally by reducing them to the same kind of observational event.

A third class of algorithms would deal with the frequencies of occurrence of the various phenomena rather than their sites of occurrence. We are coding reliability of event annotation rather than reliability of frequencies of judgment. Computation of annotation reliability based on frequencies of occurrence of particular encodings has a long history in social psychological studies of group interaction, including dialogue (Heyns and Lippitt, 19--). However, such measures are not really very relevant for our purposes. We do not base reliability judgment on frequencies because such reliability judgments would be unsuitable for demonstrating or denying the value of individual observations as data for modeling.

It is much harder to get reliability on agreement of event codings than on frequencies of the same codings. It is possible to have 100% agreement in a frequency measure and 0% agreement in an event agreement measure on the same observational data. On the other hand, 100% event agreement guarantees 100% frequency agreement as well. So the computational methods used here are much more conservative in yielding particular numerical levels of reliability than frequency methods would be on the same data.

In some of the categories it would be possible to make more recognition of partial agreement between Observers than we do, at the expense of additional complexity in the method. We have usually preferred the simpler computation, even though it tends to yield a lower score.

C. *Summary of the Methodology for Reliability Assessment*

Before going into the details of reliability assessment for each of the DAI categories, a brief recapitulation of the distinctive characteristics of the reliability assessment methodology is in order.

Reliability is computed for the annotation of individual dialogue events, rather than for relative frequencies or aggregate scores over events. A complete set of pairwise comparisons is made among Observers for each event. The Observer reliability is defined as the ratio of actual number of agreements to possible number of agreements among all pairs of Observers. No standard or correct annotation need be assumed for this method.

A distinction is made between levels of detail. Level One events are independent of all other events. The maximum number of agreements possible for each Level One annotation is N-1 for N Observers. Level Two agreement is examined only in cases where the Observer has identified a prerequisite Level One event. The maximum number of agreements possible for each Level Two annotation is M-1, where M is the number of Observers who made the prerequisite annotation.

The reliability assessment algorithm is homogeneous across annotation category. This permits aggregation of reliability scores across events and across category type. The reliability score for each event (i) is A[i]/P[i], where A is the number of agreements actually occurring and P is the maximum number possible. Aggregate reliability scores, within categories, are computed by summing the numerators and denominators for any set of events. This combined event reliability assessment can be done since each annotation is part of only one comparable reliability score. Since many alternative aggregations are possible, it is necessary to specify, before computing reliability scores, which aggregations are of theoretical or practical importance. This was done for the DAI and is reported in ***Section VII.A*** of this paper.

The major strengths of this methodology are its simplicity of category and subcategory reliability computation, its capacity to score hierarchic observations so that it fits the DAI, its use of pairwise agreement rather than comparison against "correct" or "standard" annotations, and its homogeneity across types of dialogue annotation.

The methodology has weaknesses regarding its sensitivity to the number of Observers when that number is small.

It is also sensitive to differences in observer confidence level or ambition, and it sometimes appears to magnify small differences in test-range designation so that they score wrongly as unrelated observations. The results are difficult to compare to correlation results of other studies. Ranges of text are transformed into single word units for comparison. A standard algorithm for this collapsing is described in the next section, along with detailed reliability coding rules for each category of the DAI.

V. Reliability Coding Rules by DAI Category

This section presents the reliability coding rules for each category of the DAI evaluated in this study. It is an expansion of *Section III* above in which the DAI for each category are briefly described. This section describes the steps taken to process the output of the Observers' annotation for each category in order to assess the Observer reliability. A conventional method of reducing segments or ranges of text to single words for unit comparison is utilized across categories. The category-dependent rules are specified for computing the reliability ratios for various levels of each category and for summarizing ratios across levels for each category.

A. Event Collapse Rules for Segments

All segments must first be collapsed into single words to permit comparison of unit identification among Observers. The rule for collapsing segments is to pick the main verb or, if there is no verb, noun, or if no noun, the keyword, closest to the left bracket of the reference segment. For example, [the primary word] would collapse to "word" and the previous sentence of this paragraph would collapse to "is."

Each labelled segment is treated as an individual unit for agreement assessment.

B. Requests

The output from the Requests annotation task is rather complicated since many of the annotations have lower level qualifications of fine detail. Segments are identified for both the request and response regions and also for the answer region. These segments are collapsed to unit events as described above for Repeated Reference. Level One reliability is scored with respect to identification of Requests as either a Question, Order, Directive, Rhetorical or Prohibitive. (An alternative (less conservative) computation is also made for request identification, without regard to Request type. This alternative computation is not aggregate with overall results.)

The Level Two reliability is assessed for the immediate response compliance annotation and for any eventual compliance annotations. Another Level Two reliability (with Level Two annotations as prerequisites) is computed for compliance qualification. For example, if a Question, Order or Directive is annotated as being not complied with in the response segment, a type of Non-Compliance is specified (A1-A10 or R1-R9 by the Observer.)

C. Repeated References

Observer annotations for Reference consist of labeled segments of text. Agreement is computed by counting, for each unit scored by each Observer, the number of other Observers who mark that same unit as being co-referential with at least one other common unit. The sum of these scores is the numerator for the reliability score. The reliability denominator is simply the sum of the number of units identified by each Observer, multiplied by three (the number of possible agreements for any unit marked). This score is broken down by Reference type. Separate reliability scores are computed for Text Reference, Personal (1st and 2nd person) Repeated References and Non-personal (generic You) Repeated References. The aggregate reliability score for Repeated References, Text References and Personal References is computed by adding the numerators and denominators of these separate and independent component ratios.

D. Topic

Output from the annotation task contains labelled marks of topic beginnings and endings. Each such event is collapsed into a word unit for comparison. Note that unlike other segment-annotations, those for topic are collapsed into *two* events (beginning and ending). The same rules described above for segment collapsing are used with one exception: for topic endings, the unit word is the main verb, noun or keyword nearest the right bracket. Computation of agreement considers the beginning and ending annotations as independent (rather than Level One and level two, respectively.) Thus, matches are computed for each annotation, independently of other annotations. The identification of topics already open at the start and topics still open at the end of the dialogue are counted as events comparable to any other beginning or ending of topic.

E. Expressions of Comprehension

Output from the Observer annotations includes labelled segments for comprehension expressions and comprehended regions. These segments must be collapsed into word units for comparison. The major predicate, noun or keyword nearest to the left bracket is the unit identifier.

Reliability is computed separately for four different types of comprehension expression: Positive, Negative, Selective Positive and Selective Negative Comprehension. These Level One reliabilities are combined (as with request types) to give an overall Level One Expressions of Comprehension reliability. Also, as for Requests, , an alternative (less conservative) reliability score can be computed independent of type of expression of comprehension.

Level Two annotation reliabilities are computed by counting matches on the Primary Non-Primary dimensions of qualification for each expression of comprehension identified. Level Two and Level One reliabilities are combined by adding the numerators and denominators to give an overall reliability score for Expressions of Comprehension.

F. Similar Expressions Generated Out of Context

Output from this annotation task consists of the list of "similar expression" units, generated out of context, coded first for out of context substitutability for the standard units and then (for those coded positively in the previous step) for functional substitutability in the context of the original dialogue. These annotations are subjected to Level One and Level Two reliability computations respectively. Thus, for the out of context annotations, agreement is computed among all Observers for all units.

VI. A Study of Four Dialogues - Application of the Methodology

A. Subjects

Four members of the ISI dialogue process modeling project team (the four authors) served as subjects (i.e., Observers) in the agreement test. All of the Observers had participated in the development of the Dialogue Annotation Instructions during the preceding year. Although the instructions had previously been applied to several short dialogues, this constituted the most extensive single annotation exercise for any of the Observers. The extent to which agreement could be obtained, especially among the developers of the annotation instructions, was an open issue going into this exercise. Observers were all male, native English speaking PhD graduates of American universities.

B. Dialogue Selection

Four dialogues were selected, representing two different styles of task-related, non-face-to-face, interpersonal communication. Two dialogues were excerpted from a transcript of the spacecraft - ground communications during the Apollo 13 space flight. These 10 minutes of conversation contain a total of 635 words in 66 utterances. The other two dialogues are transcripts of computer-mediated conversations between the operator of a PDP-10 computer center and two users. The operator and users are typing at terminals which are connected directly to one another by the "link" facility on the computer. This conversation was initiated by the users (referred to hereafter as the LINKERS) and contained 688 words in 80 utterances.

All dialogues received minor cosmetic treatment to correct spelling, "sanitize," and to standardize presentation format to triple spaced wide margin copy. Each sentence was numbered and each turn was labeled with the speaker's name. A replica of the transcripts presented to Observers is included as Appendix A.

C. Similar Expression Generation

A staff member of ISI, who was not one of the Observers used for annotation, was presented with a set of 146 sentences, completely out of the context in which they were uttered. These sentences were taken from the dialogues being annotated and were shuffled in order to conceal the exact context from which they came. Using the instructions on pages 46-56 of Mann *et al* (1975), this person generated similar expressions out of context for each sentence. These expressions were then retyped and formatted for presentation to the Observers. (See Appendix B)

D. Annotation of Dialogues

Each Observer, working independently in a private room, was asked to annotate all dialogues according to the DAI. Dialogues were annotated in the same order by all Observers to minimize variance within category due to fatigue and learning. All four dialogues were annotated, as a set, one category at a time. The procedure for annotation was as specified in the DAI, which is summarized in *Section III* of this paper.

Observers were granted as much time as they wanted to complete the annotations; all completed the annotation in less than 24 hours. A break between categories was permitted as long as no discussion of the annotation task took place. Actual annotation times were recorded by Observers on most categories.

The average times taken by an Observer to annotate each of the four dialogues were

Requests	12 minutes
Reference	30 minutes
Topic	5 minutes
Expressions of Comprehension	5 minutes
Similar Expressions	134 minutes

Materials used in the annotation consisted of the dialogue annotation instructions and a copy of each dialogue for each annotation category (i.e., 5 copies of each) a copy of the Similar Expressions Generated Out-of-Context, and an Observation Category Checklist (see Appendices A,B,C).

VII. Test Results and Interpretation

A. Overview of the Results

The various kinds of annotation in the DAI can be arranged in a hierarchy in more than one way, so it was necessary in setting up the test to decide what aggregate reliabilities would be computed. This was done as indicated below.

Reliability of observation is reported in Table 4 below. The indenting indicates the computation method: an item with further-indented items immediately below it is an aggregate of those items; an item with no further-indented items immediately below it is a direct independent assessment. Aggregation was performed according to the rules described in ***Section IV.B.4*** above. The composite score for Overall Reliability is a simple average of the scores for Requests, Repeated Reference, Expression of Comprehension, Topic Structure and Similar Expressions.

TABLE 4
RELIABILITY COMPUTATIONS

	OBSERVATION RATIO	OBSERVER RELIABILITY
1. Overall Reliability	(avg.)	.77
2. Reference	2437/3189	.76
3. Repeated Reference without personal pronouns	992/1557	.64
4. Text Reference	57/144	**
5. Personal pronouns	1388/1488	.93
6. Topic	526/783	.67
7. Expression of Comprehension	1682/1923	.88
8. Positive	1142/1216	.94
9. Negative	0/0	(none)
10. Selective Positive	0/0	(none)
11. Selective Negative	66/89	.74
12. Primariness		.77
13. Requests	486/659	.74
14. Questions	174/246	.71
15. Orders	4/21	**
16. Directives	12/45	**
17. Rhetoricals	0/0	(none)
18. Prohibitives	42/45	**
19. Immediate response	254/303	.84
20. Compliance	224/259	.86
21. Non-compliance Type	30/44	.68
22. Eventual compliance	26/46	.56
23. Similar Expressions	5572/6849	.82
24. Out-of-context	3362/4121	.82
25. In-context	2210/2728	.81

**There were not enough observations in this subcategory to make the reliability computation meaningful. The observations were aggregated into the category of which this subcategory is a part.

These correspond rather directly to the different kinds of annotation marks which the Observer was instructed to make, and their dependency relationships. Note that the details of the observation methods are being tested individually, but that the important general information is also available.

Another measure was computed but not included in .ne overall reliability computation, since it does not fit the overall hierarchic scheme above. It tests the reliability of identification of the fact that a Request has occurred. (In the main reliability computation for Requests, if one Observer sees a Directive and another sees an Order in the same place, these are treated as being in total disagreement. The supplementary computation described here would treats as agreeing. This allowed us to assess how much of the Request-coding unreliability was due to this kind of categorization differences.) It is reported under Requests Test Results below.

B. Requests Test Results

The overall reliability for annotation of Requests in the Four Dialogues was .74. These results represent a "very high degree" of agreement over 119 annotations identifying the five types of Request: Questions, Directives, Orders, Prohibitives and Rhetoricals (Level One) and 100 annotations of Request Compliance (Level Two). This figure indicates that three fourths of all possible pair agreements occurred.

This high reliability suggests that the phenomena of requesting and responding are fairly well explained at the structural level by the DAI There was little confusion among request types (only four events received mixed annotations.) By ignoring the disagreements with respect to request type and recomputing the overall reliability for Requests the result was .80 (rather than .74.) This alternative form of reliability computation is less conservative and would represent a relaxing of the DAI specificity. There seems to be no need to advocate such a revision of the instructions. Rather, it would improve matters simply to revise the DAI to clear up confusion between Directives and Orders (which occurred in all four mixed annotations.)

Thirty six percent (43 out of 119) of the Level One annotations were affected by "single Observer deviations." There were 16 Level One annotations with which none agreed. Four of these involved opening ceremonies which one Observer coded as Questions. Four involved annotating an answer to a request as being itself a request. Four of the deviant annotations arose from the "event collapse rule" separating parts of a single utterance. Four were simply "lone wolf" annotations (i.e. Regions annotated by only a single Observer). There were nine events on which three out of four Observers agreed. It is likely that very high, if not perfect, consensus could be obtained among these Observers through brief discussion of the rationale underlying each of these deviations. Most of the single Observer deviations seemed to be due to misinterpretation of the DAI or high sensitivity in annotation. This suggests that discussion of deviant annotations, when multiple Observers are involved with a single dialogue, may be used to increase consensus.

Forty seven percent (56 out of 119) of the Level One annotations received complete agreement from all four Observers. Twelve of these consensus events were Questions, two were Prohibitives. There were 82 Question annotations, with a reliability of .71. There were too few requests of any type other than Questions to draw conclusions about their reliability. Basically, we can conclude that the "identification" of Requests in general and questions in particular can be reliably done following the DAI.

The combined Compliance (Level Two) reliability score was .84 for the 100 annotation events which had prerequisites. The Level Two judgments on "form" of compliance/ non-compliance were understandably less reliable (.68) since they involved a forced choice from among either 4, 9 or 10 alternatives. The compliance/ non-compliance annotation was a binary choice and thus was more reliable (.86.) There were 20 annotations of Eventual Compliance, with a reliability of only .56. There were too few Repeated Request annotations to test their reliability. Compliance annotations were more reliable than Request Identifications. This is as was predicted due to the reduced choice space in the contingency annotations versus identification annotations.

The overall reliability for Request annotations (.74) is high. The only changes recommended to be made in the DAI are to clarify the instructions for distinguishing Directives and Orders. One possible extension to the DAI is the annotation of Complete Compliance (comparable to Topic Closing). This is likely to be useful in understanding dialogue, but extremely difficult to annotate and model. It is also interesting to note that in the four dialogues studied, compliance to Directives and Orders often involved merely the agreement to comply (8 annotations) rather than the desired action itself (only one annotation.) The DAI capture this distinction, but ignore complete Compliance or Non-Compliance to Requests.

Dialogue source had no apparent effect on Observer reliability for Request annotation. Although reliability scores ranged from .70 to .84, the extremes were both for Apollo dialogues and the average reliability for each *source* was identical, .74.

C. Repeated Reference Test Results

The overall reliability for Repeated Reference annotation for the four dialogues is .76. This overall result derives from a large number of highly reliable (.93) Personal Repeated Reference annotations, a small number of low reliable (.40) Text References, and a large number of moderately reliable (.64) Non-personal Repeated References.

Personal References are first and second person pronouns and personal names of the dialogue participants. Of the total of 133 occurrences of repeated personal reference, there was complete agreement among the four Observers in 77% of the cases. Most of the disagreement came from cases in which one observer failed to annotate a personal reference that the other three annotated with complete agreement (17% of the total cases). One hypothesis for these "single miss" cases is that the observer fails to see these expressions, rather than making a definite decision that these expressions are not Repeated References. This is supported informally by the surprise and chagrin of several of the Observers when questioned afterwards about their single miss cases.

Text References are expression which refer to actual text words and phrases, rather than to the concepts these words or phrases convey. There were only 24 Text Reference annotations, and the low reliability for this category can largely be attributed to the 13 "lone wolf" annotations (54% of all cases). In most of these cases in which only one observer annotated an expression as a Text Reference, the other Observers annotated the same expression as a repeated propositional reference instead. This difficulty in differentiating text references from repeated propositional references has been noted by Archbold (1975), and suggests that perhaps this distinction cannot be reliably annotated.

The Non-personal Repeated References are mostly expressions containing Non-personal pronouns or definite determiners. The reliability for the large number of these Repeated References is moderately high. Again, as for Text References, the reliability was degraded by a large number of "lone wolf" annotations. Of all the expressions marked by at least one Observer as an Non-personal Repeated Reference (209), over one third (34%) were "lone wolf" annotations. Of the three kinds of Reference being annotated, this exhibited the most variability across dialogues and across Observers.

Although there was generally considerable variation in reliability over the four dialogues (from .69 to .85), this difference wasn't due to the type of dialogue, since the two operator-linker dialogues had a combined reliability of .77 and the two Apollo dialogues .75.

There was also variation over Observers (from 0.69 to 0.83). The dominant factor here seemed to be the degree of sensitivity of the observer, since the reliability score for an observer was a decreasing function of the total number of annotations that he made.

D. Expression of Comprehension

Observers' annotations achieved very high reliability on the sub-category of Positive Comprehension (.94), weaker on both Selective Non Comprehension (.74) and overall Primariness (.77), but still very high overall for the entire category (.88). There were insufficient annotations of Negative Comprehension and Selective Positive Comprehension from which to compute reliabilities.

It seems fair to conclude that no significant change in the DAI for this category is needed.

In examining the results of the primariness annotations, an interesting pattern emerged. In the operator-linker dialogues, most comprehension was indicated implicitly (negative primariness), by about 4 to 1. In strong contrast, the Apollo dialogues exhibited a preponderance of explicit assertions of comprehension (positive primariness) by 15 to 1! This would seem to reflect the less-than-perfect communication channel used by the astronauts, as well as the pilot/military culture of the participants . (And the potential high cost of errors since the astronauts were working to save their lives during the dialogues.)

In these dialogues, Expressed Comprehension was almost always positive, with indications of some level of non-comprehension being very rare. From the obviously successful conduct of the dialogues, we can conclude that even when positive comprehension is not expressed, it is nonetheless almost always present. This suggests that, for the level of simplicity envisioned for our models, the appropriate tactic for representing the reception of an expression of positive comprehension is to do nothing, since in the absence of such expression, comprehension would have been assumed anyway.

On the other hand, the model must be sensitive to, and behave differentially in response to, expressions of non-comprehension. The very high Observer agreement on these annotations suggests that native speakers are facile in both the generation and recognition of these expressions. We anticipate that the recognition of non-comprehension, and the corresponding scope, will serve to focus the model's attention for a possible restatement, elaboration or even a correction event in the subsequent utterance.

E. Topic

Observer reliability on topic annotations (.67) was somewhat less impressive. It is encouraging to note that observations of the beginnings of tcpics are considerably more reliable than those for topic ends (by factors of from two to three), with nearly perfoct agreement on what we will (subjectively) characterize as the major topics of the dialogue. This suggests that speakers are more careful and use more definite linguistic constructions to indicate their intention to introduce a topic, and are less concerned about unambiguously terminating it. In fact, topic closings must usually be inferred by the resolution of the issues raised with the topic, rather than by anyone saying, in effect, "Let's not talk about ... anymore.". Since, in natural dialogue, issues are frequently resolved incompletely, indefinitely, or not at all, there is often no basis for being sure where a topic no longer influences the dialogue.

Besides the problems of indistinct topic endings, the other major cause for Observer disagreement was an uncertainty of the appropriate level of topic. The directions give no guidance on just how mino a topic must be to fall below the threshold of significance. So one Observer noted only the major topics, one marked just about every conceivable level of topic, with the others at arbitrary, intermediate positions. A final, lesser problem was that of the Observers simply forgetting to annotate a close for every topic that was opened.

These results lead to some tentative conclusions bearing on the revision of the DAI the scoring of the annotations, and the building of the models of the dialogue.

Some attempt should be made to give the Observer a metric for determining the appropriate level of detail for his annotations. This probably cannot be completely satisfactory since we lack any linguistic capability for precisely describing such a level (assuming we understood it with more precision). However, we can certainly make some progress over the current state of the DAI and in particular we should specifically rule out some noise-level non-topics. (e.g.: channel verification and management, and topics which begin and end in a single utterance) Some simple, coercive measures should be taken to make sure that the annotation of a topic end is a forced choice, given that it has been noted as having begun.

On the aspect of scoring, since we now score both begins and ends as Level One phenomena, we are penalizing ourselves twice for every time one observer notes a subtopic not marked by another. If we were to separate out ends as Level Two phenomena, conditional on the corresponding begins, the resulting scores would not only be "better" (higher), but would actually be more accurate. In one dialogue, with an agreement of .50 by our current methods, the Level One agreement with the proposed scoring was .64, and when combined with the Level Two was still .54.

To model the impact of topic on the conduct of a dialogue, we will have to be acutely sensitive to the forms which are used to introduce a topic as well as the body of knowledge which accompanies it. However, it would seem not to be significant were we not to be so specific about when this knowledge no longer bears on the dialogue. We imagine that a simple model of atrophy, through non-access, will suffice.

F. Similar Expressions Test Results

The reliability of Similar Expressions observation was very high for both of the kinds of judgments scored. Reliability on judging isolated expressions out-of-context was .82; reliability of judging the in-context acceptability of expressions found acceptable out of context was .81. The latter is particularly relevant to use of observations in modeling, since it indicates that judgments of the functional equivalence of two expressions taken in a particular context can be reliable.

The most frequent out-of-context annotation was "+", indicating that the given expression would be functionally equivalent to the comparison expression (from the original dialogue) under SOME circumstances. (This is a confirmation of the adequacy of the generation method, since the person who generated the similar expressions was instructed to make them functionally equivalent in this way.) However, the most frequent in-context annotation was "-", (60%), indicating that the given expression would not be functionally equivalent to the original one in THESE circumstances.

This experience with the Similar Expressions instructions indicates that they are quite adequate for their task. They yield an interesting diversity of kinds of functional non-equivalence in communication (from "-" annotations), and also an interesting diversity of kinds of changes which preserve functional equivalence of expressions (from "+" annotations).

On the other hand, we can improve the instructions for this category on the basis of this experience, particularly by changing the unit-generation and expression-generation instructions. (Long units containing embedded sentences are to be avoided. Proper names and certain other kinds of phrases require special instruction. Constraints on use of words from the original unit need to be revised.) Lower proportions of trivial cases and difficult-to-generate cases would result.

This is the only observational category for which random observation might reach interesting reliability levels. Our estimate of the reliability of a random observer generating "+", "-", and " " at the rates experienced in the test is .48 .

The reliability scoring methods are adequate, except that the whole category should be addressed on a sampling basis rather than dealing with the whole text, as was done in this test. (Over 2500 individual observations were generated in coding Similar Expressions, which all participants found excessive.)

VIII. Conclusions

A. Summary

This paper has described and demonstrated a methodology for assessing reliability for systematic observational techniques involving highly inferential, nested, content analysis of human dialogue.

This reliability assessment methodology (described in Sections IV and V) provides a conservative estimate of the Observer agreement on ***individual units*** of dialogue behavior. Most other reliability reports for systematic observational techniques only consider the ***relative frequencies*** of different annotations for different Observers on a large corpus of behaviors, for which high reliability is far easier to attain.

The method is also hierarchical, which permits the reliability assessment at successively finer levels of detail.

The reliability algorithm employs pairwise comparisons to calculate for each annotation the actual number of agreements divided by the possible number of agreements. This ratio (with numerator and denominator summation) can be computed for any level or aggregation of levels of for each category of annotation. This homogeneity greatly facilitates analysis of strengths and weaknesses of specific parts of the annotation instructions.

Despite the conservatism of the reliability assessment algorithm, very high reliability was found for the DAI. Overall reliability was .77. Dialogue category annotation reliabilities ranged from .67 to .87.

B. Interpretation of the Results

It seems important to try to understand why the DAI managed to achieve such high reliability when content analysis involving high Observer inference has notoriously poor reliability in general. Several factors which probably contributed to our high reliability are discussed in this section, followed by an interpretation of the results.

There were several characteristics of the Observers and the way in which they were trained for the annotation task which probably increased overall reliability in the present study. Observers were highly motivated and were familiar with the purpose for and eventual use of the annotations. The Observers had spent many months in debate and development of the DAI. Prior to the study reported above, a pretest was conducted on a single 150 line dialogue in order to check out the event collapse rules and the reliability computation algorithm. Discussion of disagreements and differing levels of annotation specificity probably helped to increase Observers' shared understanding of the DAI.

The DAI have several characteristics which may account for the higher reliabilities in the present study than are typical of other systematic observation techniques. First of all, the DAI make no claim to exhaustiveness. There is no theory to support such a claim with which we are familiar. Rather, the DAI focus on eclectic collection of phenomena which seem to be important for understanding how the listener in a dialogue processes information. Some utterances are annotated with respect to several observation categories, others with respect to none. The three main criteria in selecting categories for the DAI were: importance, clarity and reliability. Categories are believed to be important to the extent that communication would break down or be significantly changed in character if the phenomenon in question were omitted. Only categories for which clear instructions from which consistent annotations could be generated were included. Many predictably unreliable categories were not included in the DAI.

Reliability was enhanced by instructing Observers to annotate only clear occurrences of the phenomena, leaving out obscure cases. The results section above discussed disagreements due to one Observer annotating a marginal event. Stressing this aspect of the DAI might further increase reliability.

Finally, it should be noted that Observers were not annotating in real time. They had multiple copies of triple spaced, neatly typed transcript. It is unlikely that real-time annotation of videotapes or audio tapes would have been so reliable.

It will be important to see, in future research beyond the scope and objectives of the current project, whether Observers other than the developers of the DAI can achieve such high reliability with this instrument. Observers agreed partly to the extent that they could draw on a shared knowledge of how the English language might be used in the dialogues being analyzed. The four Observers were familiar with operator-linker dialogues, but not with Apollo Spacecraft-to-Ground communications. Yet there were no significant differences in their abilities to annotate reliably dialogues from different sources. These two facts suggest that the DAI are successfully drawing on basic, commonly used, culturally-shared knowledge about how dialogue works. This seems to be fairly independent of dialogue source. Future research can examine the extent to which other diverse sources of dialogue can be reliably annotated using the DAI. The results of the present study are most encouraging that the DAI are robust to dialogue source.

There are several reasons why the very high reliabilities found were impressive. The types of annotations required of Observers involved considerable amounts of inference. It would have been far less impressive had the DAI required lower inference annotations such as counting the number of words per turn or turns per participant, or even listing the objects or concepts referred to. In fact, most of the Observer annotations required substantial amounts of inference.

An important part of the context in which this study was conducted is our development of dialogue comprehension models, parts of which represent many of these same phenomena. The high reliability established in the present study indicates that the DAI can reliably be used to establish criteria against which to compare processes in the dialogue comprehension models. The discovery of significant structure in human dialogue, reliably disclosed by the DAI, is important to this overall research effort.

Appendix A

DIALOGUES USED IN THIS TEST

Dialogue 1

```
LINK FROM [L], JOB 20, TTY 16

L
101
Aloha /

201
anyone there? /

O
301
Yes I am  /

401
Hello /

L
501
Hi, /

601
hey I was just looking at GROUPSTAT and notice that there
are some det accounts with 48 hours piled up.  /

701
I I get det does
the system throw me out after awhile /

801
or do I just get hung on? /

O
901
I don't understand your second line, /

011
I get det does the etc. /

111
Are you asking if you detach a job will it throw you out, /
```

211
or are
you saying that when you detach a job for a certain length of
time that is it does throw you out.? /

L
311
Right, /

411
what I am asking is your second part. /

511
If I get detached,
does the system throw me out after awhile? /

O
611
No, /

711
not to my knowledge, /

811
the only way from what I understand that
you will loose that detached job is if the system happens to crash
while your job is detached./

L
911
OK. /

021
that explains the detached jobs with mucho hours piled on
it. /

121
I have been telling guys here that I thought the system did
throw you out /

221
... so I guess I will have to correct that ... well ...
misunderstanding. /

321
Thanks a lot. /

O
421
Wait, /

521
before you start correcting people let me check to be sure
that I am understanding it correctly. /

621
Because I wouldn't want to lead you wrong either. /

721
I just don't know it for a fact /

821
and I would like to get a back-up from someone who would know without
a doubt. /

921
What I will do is check on it and send you a message
or link to you later on today or first thing in the morning. /

031
So hold on for a while /

131
OK?/

L
231
Hey OK /

331
... thanks for all that. /

431
Will appreciate it. /

531
Aloha/

O
631
Aloha [operator's name]/

BREAK (LINKS)/

Dialogue 2

[CC = Capsule Communications]
[CMP = Command Module Pilot]

CC
102
Apollo 13, Houston. /

CMP
202
Go ahead. ... /

CC
302
Roger. /

402
You're coming in a little weak. /

502
Have a recommended roll rate for this PTC, if you could copy. /

CMP
602
Alright. Go ahead. /

CC
702
Okay. /

802
Recommend that you put in R1 the following: 03750 /

902
that should give you exactly a rate of 0.3 degrees per second /

012
Over. /

CMP
112
Okay. /

212
Enter 03750. /

1212
Is plus or minus our choice? /

CC
312
Roger. /

412
The same direction you rolled the last time, which I believe is plus. /

CMP
512
Okay. /

CMP
612
Hey, Vance, would you monitor our rates and kind of give an idea of when you think they're stable enough to start PTC. /

CC
712
Roger, Jack. /

812
We'll take a look and let you know as soon as they look stable enough. /

CMP
912
Okay. /

022
I've got quads A and B disabled here. /

CC
122
Roger. /

CMP
222
Have they come up with an idea of how much fuel I used on the docking and also the P23 session at 5 hours or 6 hours. /

CC
322
I think we can give you something. /

422
Stand by a minute. /

CC
522
Apollo 13, Houston. /

CMP
622
Go ahead. /

CC
722
Okay. /

822
It's looking good so far as RCS consumables are concerned, Jack. /

922
You're standing about 20 pounds above the curve right now. /

032
Looking at the TD&E, you expended 65 pounds or - Stand by - 55 pounds, correction on that. /

CMP
132
How much?

CC
232
And 14 pounds on P23s. /

1232
You used a little more out of quad A than out of the others. /

CMP
432
Okay. /

532
Thanks, Vance. /

CC
632
Roger. /

CMP
732
Hey, could you say again the TD&E fuel? /

832
We've got a different - we all heard different things. /

CC
932
I said 65 and then corrected that to 55 pounds. /

CMP
042
Okay. /

Dialogue 3

CMP	Command Module Pilot
CC	Capsule Communicator (CAP COMM)
CM	Command Module
CMC	Command Module Computer
GET	Ground Elapsed Time
LM	Lunar Module
FIDO	?

CMP
103
Joe, what are you showing for GET now?

CC
203
I think you wanted the GET, Jack, and the present GET is 96 hours 21 minutes.

303
Over.

CMP
403
Okay, thank you.

CC
503
Okay.

CC
603
And Jack, Houston.

703
For your information, FIDO tells me that we are in the Earth's sphere of influence and we're starting to accelerate.

CMP
803
I thought it was about time we crossed.

903
Thank you.

CC
013
Roger.

CMP
113
We're on our way back home.

CMP
213
There's something that puzzles me, Joe.

313
Vance mentioned yesterday that the planned entry is a CMC-guided entry, so I'm kind of curious as how are we going to get the alinement.

CC
413
Did you say how we're going to get guidance?

513
Over.

CMP
613
No.

713
How are we gong to get a platform alinement.

CC
813
Okay.

913
We got a number of interesting ideas on that

1913
and the latest one
I've heard is to power up the LM platform and aline it, and aline the CM platform to it.

CMP
023
Okay.

123
That sounds good.

CC
223
Okay.

323
And we're working out detailed procedures on that, Jack.

CMP
423
Okay.

Dialogue 4

LINK FROM [L], JOB 25, TTY 2

L:
104
Hello?

204
Would it be possible to get a scratch tape mounted for a few minutes?

O:
304
You want a tape only for a few minutes (not one that needs to be kept?)??

L:
404
Yes.

504
I'm using the MTACPY program,

1504
and I wanted to Figure out what format it writes the tape in
--I can't find any documentation on the program.

604
I have a tape here at [computer site name1]
and
I can't Figure out what format it's in.

O:
704
Have you seen a TENEX user guide??

L:
804
Yes.

904
It tells how to use the program,
but
it doesn't describe the format of the files.

014
If it's not possible, I can understand.

O:
114
It will take a minute..

214
Please stand by...

[operator checks which tape units are available]

O:
314
Use MTA1

L:
414
OK, Thanks.

514
Also,
I was wondering, I want to mail out some tapes that I have here.

614
To whom do I address them (and how do I identify them)?

O:
714
USC-ISI, 4676 Admiralty Way, Marina del Rey, CA. 90291, c/o [name1].

814
Please identify with [computer site name1] tape

914
Also,
can the [computer site name1] account use all of them at any time

1914
(i.e. what is the restriction list)

L:
024
Hmm...

124
I don't know--

1124
I didn't know there was one.

O:
224
This is just a list saying who may use those tapes--
the operator will have to look up in the list to see if a user may use the tape..

324
If you're to be the only one, fine...

L:
424
Yes,

2424
we'll probably be the only people using them,
but
I suppose that we can send that along with the tapes (?)
Is it easier if we restrict usage to ourselves?

O:
524
It might be,

2524
but if you need other accounts to be able to write on them, we'll have to be told..

1524
We are not really tape oriented here, so we have to put some of the burden on users as to whom may play with their tapes..

L:
824
I see.

924
Well,
we won't be using them for too long,

2924
we expect to get our system up in a month-or-so,

3924
and we'll be on the net.

034
So....

O:
134
Fine...

234
Just send tapes with appropriate labels then

L:
334
OK,

434
Thanks a lot --

1434
I'll let you know when I'm done with the tape.

O:
534
Thanks.

634
Bye.

L:
734
Bye

O:
BREAK

Appendix B

SAMPLE SIMILAR EXPRESSIONS

The similar expressions generated for this test are shown below on the right for one of the dialogues, with the original dialogue shown on the left. The numbers identify the units generation. Units for which no similar expressions appear are duplicates whose expressions were generated elsewhere.

CMP
103
Joe, what are you showing for GET now?

CC
203
I think you wanted the GET, Jack, and the present GET is 96 hours 21 minutes.

303
Over.

CMP
403
Okay, thank you

CC
503
Okay.

CC
603
And Jack, Houston.

703
For your information, FIDO tells me that we are in the Earth's sphere of influence and we're starting to accelerate.

CMP
803
I thought it was about time we crossed.

903
Thank you.

CC
013
Roger.

CMP
113
We're on our way back home.

103

1. Joe, what's the number of the GET dial?
2. Hey you.
3. Mr. Black, what styles are you showing for spring now?

203

1. You wanted to know how long you've been out and the answer is 96 hours, 21 minutes.
2. It takes 96 hours, 21 minutes to get to the moon, Jack.
3. The Greatest Eating Time is 96 hours, 21 minutes.

403

1. Right, thanks.
2. Fine, I thank you.
3. A-okay.

603

1. And Tom Mix.
2. And John Houston.
3. And him.
4. And them.

703

1. If you'd like to know, my fortune teller says we are in the earth's sphere of influence but moving toward another.
2. He tells me we're influenced by the earth but soon we'll be moving on to be influenced by a new planet.
3. We're still tied to the earth but pulling away slowly.

803

1. It's time we met.
2. It's time to intersect lines.
3. Now we should try the hybrid.

113

1. We're coming home.
2. We're going to our house.
3. We'll soon be at our apartment.

CMP
213
There's something that puzzles me, Joe.

313
Vance mentioned yesterday that the planned entry is a CMC-guided entry, so I'm kind of curious as how are we going to get the alinement.

CC
413
Did you say how we're going to get guidance?

513
Over.

CMP
613
No.

713
How are we gong to get a platform alinement.

CC
813
Okay.

913
We got a number of interesting ideas on that

1913
and the latest one
I've heard is to power up the LM platform and aline it, and aline the CM platform to it.

CMP
023
Okay.

123
That sounds good.

CC
223
Okay.

323
And we're working out detailed procedures on that, Jack.

CMP
423
Okay.

213
1. I'm bewildered by something, dear.
2. I don't completely understand that, pal.
3. It confuses me, buddy.

313
1. He's told me that arrival to be of the CMC-guided type, so I want to know how we're to get it arranged.
2. How will we ever get everything arranged when arrival is to be that special guided type?
3. He told us yesterday the intentional arrival will be of the CMC type, so how will we get the arrangements made?

413
1. Do you know in which way we will obtain advice?
2. From whom will we get directions?
3. How will we get the instructions?

513
1. Finished.
2. Beyond.
3. Recovored.

613
1. I can't.
2. I'd love to but....
3. Absolutely not.

713
1. Will we reach agreement on a political policy statement?
2. How will we get policy affiliation?
3. How will we get the stage arranged?

913
1. People contributed stimulating opinions on that particular subject.
2. There were many provocative thoughts brought forth
3. Several attractive notions were offered

1913
1. Beef up the first stage and tie-in, then tie-in the second stage to it
2. Strengthen the first policy statement and get an alliance, then tie-in the second policy statement.

123
1. That's cool.
2. The music is beautiful.
3. It's OK with me.

323
1. We're developing policies in that area, Jack
2. We will formulate meticulous methods for that, Jack
3. We're getting down to the nitty gritty.

Appendix C

CHECKLIST OF DIALOGUE ANNOTATION TASKS FOR OBSERVERS

A. Repeated Reference

1. Identify Repeated References

    a. underline reference phrases
    b. label with a common number
    c. overline embedded reference phrases and label
    d. for pronouns:
        1) underline (but don't label) singular 1st and 2nd-person pronouns
        2) label plural 1st and 2nd-person pronouns
        3) circle (but don't label) Non-personal 2nd-person pronouns
        4) distinguish possessor and possessed for pronominal possessives
    e. do not annotate sets/subsets/elements or treat the latter as co-referential with the former

2. Identify text references

    a. underline text references and the text referred to
    b. label with "TR" and a common number

B. Requests

1. Identify questions (immediate, specific, verbal response)

    a. delimit question phrase with angle brackets
    b. label phrase and immediately following turn
    c. delimit response phrase(s), if any, in following turn with double angle brackets << >>
    d. mark response phrase for compliance (+, -)
    e. if response is non-compliant, qualify with "A1-A10"
    f. go back over transcript and for each question:
        1) delimit answer region (general segment markers), if any
        2) label answer region "partial" if appropriate
        3) label answer region to distinguish different views on when or whether an answer was given

2. Identify orders (immediate, specific, nonverbal behavior)

a. delimit order phrase with angle brackets
b. label phrase and immediately following turn
c. delimit response phrase(s), if any, in following turn with double angle brackets << >>
d. mark response phrase for compliance (+, -)
e. if response is compliant, qualify with "C1-C3"
f. if response is non-compliant, qualify with "R1-R9"
g. go back over the transcript and for each order:
  1) identify any response region (other than the already delimited <<immediate response>>) with general segment markers
  2) label response region "partial" if appropriate
  3) label response region to distinguish different views on when or whether compliance was made

3. Identify directives (non-immediate, verbal or nonverbal behavior)

a. delimit directive phrase with angle brackets
b. label phrase and immediately following turn
c. delimit response phrase(s), if any, in following turn with double angle brackets << >>
d. mark response phrase for compliance (+, -)
e. if response is compliant, qualify with "C1-C3"
f. if response is non-compliant, qualify with "A1-A10"
g. go back over the transcript and for each directive:
  1) identify any response region (other than the already delimited <<immediate response>> with general segment markers
  2) label response region "partial" if appropriate
  3) label response region to distinguish different views on when or whether compliance was made

4. Identify Rhetoricals and Prohibitives

a. delimit the phrase comprising the rhetorical or prohibitive with angle brackets
b. label with R or P respectively
c. do not annotate the "following turn" as for questions
d. go back over the transcript and for each R and P identify occurrence of the unexpected behavior:
  1) delimit these with the general segment markers
  2) label them with the corresponding label

5. Identify misunderstandings

a. denote any passage which indicates a misunderstanding
b. summarize in your own words its nature

6. For repeated requests

a. label repetitions of requests with an " = " prefix

7. Watch out for pseudo requests

a. statements which describe a behavior but do not create an expectation or commitment to respond (e.g., those for which no response is given) should not be annotated as requests

C. Expression of Comprehension

1. Identify positive comprehension

a. delimit explicit and implicit expressions of positive comprehension with angle brackets

b. label with "PC"

c. identify a region for which comprehension is expressed

1) if preceding turn, add a " / " to the label

2) if other than preceding turn, delimit with general segment markers and corresponding label

d. if degree of comprehension is indefinite, add "P1" to the expression label (otherwise "P2" is assumed)

2. Identify noncomprehension

a. delimit explicit and implicit indications of noncomprehension with angle brackets

b. label with "NC"

c. identify the region not comprehended

1) if preceding turn, add a " / " to the label

2) if other than preceding turn, delimit with general segment markers and corresponding label

d. if degree of noncomprehension is indefinite, add "N1" to the expression label (otherwise "N2" is assumed)

3. Identify selective comprehension

a. delimit explicit and implicit indications of selective comprehension with angle brackets

b. label with "SPC" or "SNC"

c. identify the region indicated and delimit with the general segment markers and corresponding label

d. if degree of comprehension is indefinite, add "P1" or "N1" to the expression label (otherwise "P2" or "N2" is assumed)

4. Distinguish primary/nonprimary expressions of comprehension

   a. if the expression region communicates primarily (i.e., only or mostly) comprehension or noncomprehension, add "++" to the label
   b. if the expression region definitely communicates additional information (e.g., agreement, approval, consent, answer to a request), add "--" to the label

D. Topic Structure

1. Identify distinct topics

   a. delimit the utterance with which each distinct topic begins and ends for each speaker with general segment markers
   b. label each beginning and ending with a brief title (speaker A's labels in the left margin, B's in the right)
   c. use the same label if a topic reopens or is shared by the two speakers
   d. go back over the transcript and list any topics that were already open at the start or still not closed at the end

Appendix D

A Procedural Specification of the Agreement Computation

The algorithm below expresses the general part of the agreement computation. The process language is intended to be "Algol-like," readable by people who know any of the languages in the Algol family, but with many of the obvious programming necessities left out for readability.

The algorithm has 3 arrays for holding the running agreement count, the running possible agreement count, and the final ratio. Each of these arrays is one dimensional with a length equal to the number of different recognized event types.

It consists of a main body and several supplementary procedures whose function is described in the table below.

NAME	FUNCTION
EVENT-AT(PLACE)	DETERMINES WHETHER THERE IS AN EVENT IN ANY EVENT STREAM AT A GIVEN PLACE.
EVENT(PLACE,N)	DETERMINES WHETHER THERE IS AN EVENT IN A PARTICULAR EVENT STREAM AT A GIVEN PLACE.
PAIRAGREE(PLACE,FIRS' GUY,SECONDGUY)	DECIDES WHETHER 2 EVENTS AT A PLACE AGREE
AGREE-COUNTER(PLACE,N)	COUNTS THE NUMBER OF ACTUAL AGREEMENTS WITH A PARTICULAR EVENT.
POSSIBLECOUNTER(PLACE,N)	COUNTS THE NUMBER OF POSSIBLE AGREEMENTS WITH A PARTICULAR EVENT.
PREREQUISITES(PLACE,TYPE,N)	DECIDES WHETHER AT A PARTICULAR PLACE IN A PARTICULAR EVENT STREAM, THE POSSIBILITY PREREQUISITES FOR A PARTICULAR EVENT TYPE ARE SATISFIED.
EVENTTYPE(PLACE,INDEX)	YIELDS THE TYPE OF A PARTICULAR EVENT. TYPE ENCODES ALL OF THE NECESSARY EVENT PROPERTY INFORMATION, SO THAT EVENTS AGREE IFF THEIR TYPES ARE EQUAL.

Initialization sets the following values: TEXTSIZE, OBSERVER-COUNT, TYPECOUNT, ILLFORM ← 0.

MAIN BODY

```
 BEGIN
FOR PLACE ← 1 STEP 1 UNTIL TEXTSIZE DO

IF EVENT-AT(PLACE) THEN

 BEGIN
           FOR INDEX ← 1 STEP 1 UNTIL OBSERVER-COUNT DO

           IF EVENT(PLACE,INDEX) THEN
 BEGIN
            AGREE-COUNTER(PLACE,INDEX);
           POSSIBLECOUNTER(PLACE,INDEX)
 END;
           RATIOCOMPUTE()
 END;
```

SUPPLEMENTARY PROCEDURES

```
PROCEDURE AGREE-COUNTER(SPOT,OBSERVER-INDEX)
 BEGIN

           CASETYPE ← EVENTTYPE(SPOT,OBSERVER-INDEX);

           IF NOT PREREQUISITES(SPOT,OBSERVER-INDEX,CASETYPE) THEN
 BEGIN
                     INCREMENT(ILLFORM);

COMMENT: BY DEFINITION, ONE CANNOT AGREE WITH ILLFORMED ANNOTATIONS;
 END
           ELSE
           FOR I ← 1 STEP 1 UNTIL N DO

IF NOT (N=OBSERVER-INDEX) THEN
 BEGIN
           IF PAIRAGREE(SPOT,OBSERVER-INDEX,I)

           THEN INCREMENT(AGREE-COUNT[CASETYPE]);
 END;
 END;
```

```
PROCEDURE POSSIBLECOUNTER(SPOT,OBSERVER-INDEX)
 BEGIN
          CASETYPE ← EVENTTYPE(SPOT,OBSERVER-INDEX);

          FOR I ← 1 STEP 1 UNTIL OBSERVERCOUNT DO

          IF
(NOT (I = OBSERVER-INDEX)) AND PREREQUISITES(SPOT, OBSERVER-INDEX, I)
          THEN

          INCREMENT(POSSIBLECOUNT[CASETYPE]);
 END;


PROCEDURE RATIOCOMPUTE();
FOR TYPE ← 1 STEP 1 UNTIL TYPECOUNT DO

TYPESCORE[TYPE] ← IF POSSIBLE-COUNT[TYPE] = 0 THEN 1 ELSE
          AGREE-COUNT[TYPE] / POSSIBLE-COUNT[TYPE]


PROCEDURE INCREMENT(COUNT):
          COUNT ← COUNT + 1;
```

The PAIRAGREE and PREREQUISITES procedures are observation-category dependent, and so are not described here.

References

Archbold, A. A., "Text Reference And Repeated Propositional Reference: Concepts And Detection Procedures", *Working Papers in Dialogue Modeling - Vol I*, USC/Information Sciences Institute, ISI/RR-77-55, January 1977, Sec. 2.

Bales, Robert F., *Interaction Process Analysis*, Addison-Wesley, Cambridge, Mass., 1951.

Heyns, Roger W., and Ronald Lippitt, "Systematic Observational Techniques", in Gardner Lindzey (Ed.), *Handbook of Social Psychology, Volume I, Theory and Method*, Addison-Wesley, Reading, Mass., 1954, pp. 370-404.

Mann, W. C., *Dialogue-Based Research in Man-Machine Communication*, USC/Information Sciences Institute, ISI/RR-75-41, November 1975.

Mann, W. C., J. A. Moore, J. A. Levin, and J. H. Carlisle, *Observation Methods For Human Dialogue*, USC/Information Sciences Institute, ISI/RR-75-33, June 1975.

Newell, A., "On The Analysis of Human Problem Solving Protocols", J. C. Gardin and B. Jaulin, *Calcul et Formalisation dans les Sciences de L'Homme*, Centre National de la Recherche Scientifique, Paris, 1968.

Newell, A. and Simon, H. A. *Human Problem Solving*, Prentice-Hall, Englewood Cliffs, New Jersey, 1972.

Waterman, D. A., and A. Newell, "PAS-II: An Interactive Task-Free Version of An Automatic Protocol Analysis System", Department of Computer Science, Carnegie-Mellon University, June, 1973.

DISTRIBUTION LIST

AC-S, Education Programs
Education Center, MCDEC
Quantico, VA 22134

AFHRL-AS (Dr. G.A. Eckstrand)
Wright-Patterson AFB
Ohio 45433

ARI Field Unit - Leavenworth
P.O.Box 3122
Ft. Leavenworth, KS 66027

Advanced Research Projects Agency
Administrative Services
1400 Wilson Blvd.
Arlington, VA 22209
Attn: Ardella Holloway

Air University Library
AUL-LSE 76-443
Maxwell AFB, AL 36112

Prof. Earl A. Alluisi
Code 287
Dept. of Psychology
Old Dominion University
Norfolk, VA 23508

Dr. Daniel Alpert
Computer-Based Enducation
Research Laboratory
University of Illinois
Urbana, IL 61801

Dr. John R. Anderson
Dept. of Psychology
Yale University
New Haven, CT 06520

Armed Forces Staff College
Norfolk, VA 23511
Attn: Library

Ms. Carole A. Bagley
Applications Analyst
Minnesota Educational
Computing Consortium
1925 Sather Ave.
Lauderdale, MN 55113

Dr. James Baker
U.S. Army Research Institute
1300 Wilson Blvd.
Arlington, VA 22209

Dr. M.A. Bertin, Scientific Director
Office of Naval Research
Scientific Liaison Group-Tokyo
American Embassy
APO San Francisco 96503

Dr. Jack R. Borsting
Provost & Academic Dean
U.S. Naval Postgraduate School
Monterey, CA 93940

Dr. John Brackett
SofTech
460 Totten Pond Rd.
Waltham, MA 02154

Dr. Robert K. Branson
1A Tully Bldg.
Florida State University
Tallahassee, FL 32306

Dr. John Seeley Brown
Bolt, Beranek, and Newman, Inc.
50 Moulton St.
Cambridge, MA 02138

Dr. Victor Bunderson
Institute for Computer Uses
in Education
355 EDLC Brigham Young University
Provo, UT 84601

Dr. Ronald P. Carver
School of Education
University of Missouri-Kansas City
5100 Rockhill Rd.
Kansas City, MO 64110

Jacklyn Caselli
ERIC Clearinghouse on
Information Resources
Stanford University
School of Education - SCRDT
Stanford, CA 94305

Century Research Corporation
4113 Lee Highway
Arlington, VA 22207

Chairman, Leadership & Law Dept.
Div. of Professional Development
U.S. Naval Academy
Annapolis, MD 21402

Chief of Naval Education and
Training Support (01A)
Pensacola, FL 32509

Dr. Kenneth E. Clark
College of Arts & Sciences
University of Rochester
River Campus Station
Rochester, NY 14627

Dr. Allan M. Collins
Bolt, Beranek and Newman, Inc.
50 Moulton St.
Cambridge, MA 02138

Dr. John J. Collins
Essex Corporation
6305 Caminito Estrellado
San Diego, CA 92120

Commandant
U.S. Army Institute of Administration
Attn: EA
Fort Benjamin Harrison, IN 46216

Commanding Officer
Naval Research Laboratory
Code 2627
Washington, DC 20390

Commanding Officer
Naval Health Research Center
San Diego, CA 92152
Attn: Library

Capt. H.J. Connery, USN
Navy Medical R&D Command
NNMC, Bethesda, MD 20014

Dr. T.E. Cotterman
AFHRL-ASR
Wright Patterson AFB
Ohio 45433

Mr. Joseph J. Cowan, Chief
Psychological Research Branch
(G-P-1-62)
U.S. Coast Guard Headquarters
Washington, DC 20590

DCDR, USAADMINCEN
Bldg. #1, A310
Attn: AT21-OED Library
Ft. Benjamin Harrison, IN 46216

Dr. Ruth Day
Dept. of Psychology
Yale University
2 Hillhouse Ave.
New Haven, CT 06520

Defense Documentation Center
Cameron Station, Bldg. 5
Alexandria, VA 22314
Attn: TC

Director, Office of Manpower
Utilization
HQ, Marine Corps (Code MPU)
BCB, Building 2009
Quantico, VA 22134

Director, Management Information
Systems Office
OSD, M&RA
Room 3B917, The Pentagon
Washington, DC 20301

Dr. Donald Dansereau
Dept. of Psychology
Texas Christian University
Fort Worth, TX 76129

Dr. Ralph Dusek
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

ERIC Facility-Acquisitions
4833 Rugby Ave.
Bethesda, MD 20014

Dr. John Eschenbrenner
McDonnell Douglas Astronautics
Company-East
PO Box 30204
St. Louis, MO 80230

Major I. N. Evonic
Canadian Forces Personnel
Applied Research Unit
1107 Avenue Road
Toronto, Ontario, CANADA

Dr. Beatrice Farr
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

Dr. Marshall J. Farr, Director (4)
Personnel & Training Research Programs
Office of Navy Research (Code 458)
Arlington, VA 22217

Dr. James M. Ferstl
Employee Development:
Training Technologist
Bureau of Training
U.S. Civil Service Commission
Washington, DC 20415

Dr. Victor Fields
Dept. of Psychology
Montgomery College
Rockville, MD 20850

Dr. John Ford
Navy Personnel R&D Center
San Diego, CA 92152

Dr. Larry Francis
University of Illinois
Computer Based Educational
Research Laboratory
Champaign, Il 61801

Dr. John R. Frederiksen
Bolt, Beranek & Newman, Inc.
50 Moulton St.
Cambridge, MA 02138

Dr. Alfred R. Fregly
AFOSR-NL, Building 410
Bolling AFB, DC 20332

Dr. Frederick C. Frick
MIT Lincoln Laboratory
Room D 268
PO Box 73
Lexington, MA 02173

Dr. Vernon S. Gerlach
College of Education
146 Payne Bldg. B
Arizona State University
Tempe, AZ 85281

Dr. Robert Glaser
Co-Director
University of Pittsburgh
3939 O'Hara St.
Pittsburgh, PA 15213

Capt. D.M. Gragg, MC, USN
Head, Section on Medical Education
Uniformed Services Univ. of
the Health Sciences
6917 Arlington Road
Bethesda, MD 20014

Dr. Duncan Hansen
School of Education
Memphis State University
Memphis, TN 38118

Dr. Frank J. Harris
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

Col. Frank Hart, Director
Training Management Institute
U.S. Army, Bldg. 1725
Fort Eustis, VA 23604

Col. G.B. Howard
U.S. Army
Training Support Activity
Fort Eustis, VA 23604

HumRRO-Columbus Office
Suite 23
2601 Cross Country Dr.
Columbus, GA 31906

HumRRO-Ft. Knox Office
P.O. Box 293
Fort Knox, KY 40121

HumRRO-Western Division
27857 Berwick Drive
Carmel, CA 93921
ATTN: Library

Human Resources Research Organization
400 Plaza Bldg.
Pace Blvd. at Fairfield Drive
Pensacola, FL 32505

Dr. Earl Hunt
Dept. of Psychology
University of Washington
Seattle, WA 98105

Instructional Technology Branch
AFHRL
Lowry AFB, CO 80230

Dr. Edgar Johnson
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

Dr. Wilson A. Judd
McDonnell-Douglas Astronautics Co.
East
Lowry AFB
Denver, CO 80230

Dr. Arnold F. Kanarick
Honeywell, Inc.
2600 Ridgeway Pkwy.
Minneapolis, MN 55413

Dr. Milton S. Katz, Chief
Individual Training & Performance
Evaluation Technical Area
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

Dr. Steven W. Keele
Dept. of Psychology
University of Oregon
Eugene, OR 97403

Dr. David Klahr
Dept. of Psychology
Carnegie-Mellon University
Pittsburgh, PA 15213

Dr. Alma E. Lantz
University of Denver
Denver Research Institute
Industrial Economics Division
Denver, CO 80210

Dr. Sylvia R. Mayer (MCIT)
HQ Electronic System Division
LG Hanscom Field
Bedford, MA 01730

William J. McLaurin
Room 301
Internal Revenue Service
2221 Jefferson Davis Hwy.
Arlington, VA 22202

Dr. Donald E. Meyer
U.S. Air Force
ATC-XPTD
Randolph AFB, TX 78148

Military Assistant for Human Resources
Office of the Director of Defense
Research & Engineering
Room 3D129, The Pentagon
Washington, DC 20301

Dr. Andrew R. Molnar
Science Education Dev. & Res.
National Science Foundation
Washington, DC 20550

Dr. Ross L. Morgan (AFHRL-ASR)
Wright-Patterson AFB
Ohio 45433

Dr. Leo Munday
Houghton Mifflin Co.
P.O. Box 1970
Iowa City, IA 52240

Navy Personnel R&D Center
Code 306
San Diego, CA 92152
Attn: Dr. James McGrath

Navy Personnel R&D Center
San Diego, CA 92152
Attn: Dr. J.D. Fletcher

Navy Personnel R&D Center
Code 1
San Diego, CA 92152
Attn: Dr. James McGrath

Navy Personnel R&D Center
San Diego, CA 92152
Attn: Library

Dr. Leon Nawrocki
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

Dr. Donald A. Norman
Dept. of Psychology C-009
University of California, San Diego
La Jolla, CA 92093

Dr. Harold F. O'Neil, Jr.
Advanced Research Projects Agency
Cybernetics Technology, Room 623
1400 Wilson Blvd.
Arlington, VA 22209

Mr. Thomas C. O'Sullivan
TRAC
1220 Sunset Plaza Dr.
Los Angeles, CA 90069

ONR Branch Office
495 Summer Street
Boston, MA 2210
Attn: Dr. James Lester

ONR Branch Office
1030 East Green Street
Pasadena, CA 91101
Attn: Dr. Eugene Gloye

ONR Branch Office
536 S. Clark Street
Chicago, IL 60605
Attn: Dr. Charles E. Davis

Office of Naval Research
Code 200
Arlington, VA 22217

Mr. Luigi Petrullo
2431 N. Edgewood St.,
Arlington, VA 22207

Dr. Kenneth A. Polycyn
PCR Information Sciences Co.
Communication Satellite Applications
7600 Old Springhouse Rd.
McLean, VA 22101

Principal Civilian Advisor
for Education and Training
Naval Training Command, Code 00A
Pensacola, FL 32508
Attn: Dr. William L. Maloy

R.Dir. M. Rauch
P II 4
Bundesministerium der Verteidigung
Postfach 161
53 Bonn 1, GERMANY

Research Branch
AFMPC-DPMYP
Randolph AFB, TX 78148

Dr. Joseph W. Rigney
University of So. California
Behavioral Technology Laboratories
3717 S. Grand
Los Angeles, CA 90007

Dr. Marty Rockway (AFHRL-TT)
Lowry AFB
Colorado 80230

Dr. Andrew M. Rose
American Institutes for Research 1055
Thomas Jefferson St. NW
Washington, DC 20007

Dr. Leonard L. Rosenbaum
Chairman
Dept. of Psychology
Montgomery College
Rockville, MD 20850

Dr. Worth Scanland
Chief of Naval Education & Training
NAS, Pensacola, FL 32508

Scientific Advisor to the Chief
of Naval Personnel (Pers Or)
Naval Bureau of Personnel
Room 4410, Arlington Annex
Washington, DC 20370

Dr. Robert J. Seidel
Instructional Technology Group, HumRRO
300 N. Washington St.
Alexandria, VA 22314

A.A. Sjoholm, Head Technical Support
Navy Personnel R&D Center
Code 201
San Diego, CA 92152

Dr. A. L. Slafkosky
Scientific Advisor (Code RD-1)
HQ, U.S. Marine Corps
Washington, DC 20380

Dr. Marshall S. Smith
Associate Director
NIE-OPEPA
National Institute of Education
Washington, DC 20208

Dr. Alfred F. Smode, Director
Training Analysis & Evaluation Group
Department of the Navy
Orlando, FL 32813

Dr. Richard Snow
Stanford University
School of Education
Stanford, CA 94305

LCDR J.W. Snyder, Jr.
F-14 Training Model Manager
VF-124
San Diego, CA 92025

Dr. William Strobie
McDonnell-Douglas Astronautics Co.
East
Lowry AFB
Denver, CO 80230

Dr. Persis Sturgis
Dept. of Psychology
California State University-Chico
Chico, CA 95926

Superintendent (Code 1424)
Naval Postgraduate School
Monterey, CA 93940

Technical Director
U.S. Army Research Institute for the
Behavioral & Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

Capt. Jack Thorpe, USAF
AFHRL-FTS
Williams AFB, AZ 85224

Mr. Walt W. Tornow
Control Data Corporation
Corporate Personnel Research
P.O. Box 0-HQN060
Minneapolis, MN 55440

Dr. Benton J. Underwood
Dept. of Psychology
Northwestern University
Evanston, IL 60201

Dr. Carl R. Vest
Battelle Memorial Institute
Washington Operations
2030 M Street NW
Washington, DC 20036

Dr. Joseph Ward
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

Dr. Claire E. Weinstein
Educational Psychology Dept.
University of Texas at Austin
Austin, TX 78712

Dr. David J. Weiss
Dept. of Psychology
N660 Elliott Hall
University of Minnesota
Minneapolis, MN 55455

Dr. Keith Wescourt
Dept. of Psychology
Stanford University
Stanford, CA 94305

Dr. Joseph L. Young
Director
Memory & Cognitive Processes
National Science Foundation
Washington, DC 20550

Robert Young
Advanced Research Projects Agency
1400 Wilson Blvd.
Arlington, VA 22209

USC / INFORMATION SCIENCES INSTITUTE
4676 Admiralty Way Marina del Rey California 90291